DOCUMENT RESUME

ED 476 138 TM 034 890

AUTHOR	DeMars, Christine E.
TITLE	Recovery of Graded Response and Partial Credit Parameters in MULTILOG and PARSCALE.
PUB DATE	2002-04-00
NOTE	28p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 21-25, 2003).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS	*Computer Software; Item Response Theory; Simulation; *Statistical Analysis; Statistical Distributions
IDENTIFIERS	Graded Response Model; *MULTILOG Computer Program; *PARSCALE Computer Program; Partial Credit Model

ABSTRACT

Using simulated data, the MULTILOG and PARSCALE software packages were compared for their recovery of item and trait parameters under the graded response and generalized partial credit item response theory models. The shape of the latent population distribution (normal, skewed, or uniform) and the sample size (250 or 500) were varied. Parameter estimates were essentially unbiased under all conditions, and the root mean square error was similar for both software packages. The choice between these packages can therefore be based on considerations other than the accuracy of parameter estimation. (Contains 3 tables, 5 figures, and 22 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 476138

Running head: MULTILOG and PARSCALE

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

C. DeMars

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Recovery of Graded Response and Partial Credit Parameters in MULTILOG and PARSCALE

Christine E. DeMars

James Madison University

TM034890

Paper presented at the annual meeting of the American Educational Research Association, Chicago. (2002, April).

BEST COPY AVAILABLE

Abstract

Using simulated data, MULTILOG and PARSCALE were compared on their recovery of item and trait parameters under the graded response and generalized partial credit item response theory models. The shape of the latent population distribution (normal, skewed, or uniform) and the sample size (250 or 500) were varied. Parameter estimates were essentially unbiased under all conditions, and the root mean square error was similar for both software packages. The choice between these packages can therefore be based on considerations other than the accuracy of parameter estimation.

Recovery of Graded Response and Partial Credit Parameters in MULTILOG and PARSCALE

MULTILOG (Thissen, 1991) and PARSCALE (Muraki & Bock, 1997) are two commercially-available software packages that will provide item parameter and trait parameter estimates for a variety of polytomous models. Both packages will estimate Samejima's (1969) graded response model, Masters' (1982) partial credit model (including a generalized partial credit model, an extension that allows different slopes across items), and the 1, 2, and 3-parameter logistic models. In addition, PARSCALE can be used for Andrich's (1978) rating scale model (and a variant with unequal slopes, as well as a rating-scale analogue for the graded response model). MULTILOG can be used for the nominal response model (Bock, 1972) and the multiple-choice model (Thissen & Steinberg, 1984). Both products use marginal maximum likelihood, with a series of quadrature points approximating the density at discrete points of the latent population distribution. In PARSCALE, either the normal distribution can be assumed for the latent distribution, or the shape of the distribution can be approximated by estimating the density at each quadrature point after each iteration in the item parameter estimation (scaling it after each step to have a mean of zero and standard deviation of one, but not necessarily a normal distribution). In MULTILOG, though the metric of the item parameters is still scaled such that the mean of the estimated latent distribution is zero and the standard deviation is one, the normal distribution is assumed unless the user requests estimation of the population distribution with Johnson curves, a complicated procedure "not recommended for routine or casual use" (Thissen, 1991, p. C-1).

Though there have been studies comparing software packages for dichotomous items (Carlson & Locklin, 1995; Drasgow, 1989; Kirisci, Hsu, & Yu, 2001; Mislevy & Stocking, 1989; Ree, 1979; Swaminathan & Gifford, 1983; Yen, 1987), there has been little work comparing IRT packages for polytomous items. Childs and Chen (1999) illustrated how the parameters from

MULTILOG and PARSCALE could be put on the same metric. Using the graded response model and the generalized partial credit model, they gave an example based on a single set of real data. They showed that MULTILOG and PARSCALE provided similar item parameter estimates in their dataset (differences ranged from 0.00 to 0.06 for the *a* parameters and 0.00 to 0.08 for the *c* parameters), but only a single dataset was studied.

Several researchers have examined the recovery of parameters in MULTILOG alone. Reise and Yu (1990), working with the graded response model, found discrimination and category RMSE was about the same for normal and skewed population distributions, and slightly smaller for uniform distributions (and conversely for the correlation between true and estimated parameters). Sample size made the largest difference in RMSE for the item parameters (the authors suggested a minimum sample size of 500 as a general heuristic). For the ability parameters, estimated through modal a-posterior (MAP) methods with a normal prior, the RMSE was slightly, but not substantially, larger for the uniform distribution.

Choi, Cook, and Dodd (1997) studied the recovery of partial credit model item and ability parameters (the traditional model with equal slopes, not the generalized partial credit model). They varied sample size, number of items, and number of item categories, and found that the sample size needed to have reasonable correlations and RMSEs between estimated and true category parameters depended on the number of item categories. When there were more categories, the RMSEs for item parameters were larger. The accuracy of the ability estimates was influenced more by the number of items and the number of categories (increases in either led to improved estimation) than the sample size used to calibrate the items.

Some of the work with dichotomous items showed non-normal trait distributions led to poorer estimates of the item parameters, particularly discrimination (Ree, 1979; Stone, 1992; Swaminathan & Gifford, 1983). In Swaminathan & Gifford, skewed distributions were more

problematic than uniform or platykurtic distributions. Ree found lower correlations between true and estimated parameters, especially discrimination and guessing, for a skewed distribution than for uniform or normal distributions. Stone looked at item and ability recovery with the two-parameter model using MULTILOG. He found the non-normal distributions led to greater bias in item discrimination estimates, but did not greatly effect RMSE of discrimination or bias/RMSE for item difficulty or ability. Seong (1990), using BILOG and a 2-PL model, varied both the prior distribution and the data distribution. Both item difficulty and discrimination were estimated somewhat better (smaller bias and RMSE) when the prior matched the data; ability estimates were influenced even more, but Seong used EAP estimation so the ability parameters were directly affected by the prior distribution, not just through the effect of the prior on the item parameter estimates. In contrast, Kirisci, Hsu, and Yu (2001), found little effect for distribution shape on estimating either item or person parameters. Similarly, Reise and Yu (1990) found RMSE was equal for normal and skewed distributions, and only slightly smaller for uniform distributions.

The purpose of this study was to compare MULTILOG and PARSCALE on accuracy in item and person parameter recovery for the graded response and generalized partial credit models. Because both programs use marginal maximum likelihood estimation (though the exact algorithms may differ), it was expected that the results would generally be similar for both programs, except possibly when the data were drawn from a non-normal distribution (because MULTILOG does not adjust the estimated latent population distribution beyond the first two moments).

Method

Data simulation

Two models were studied, the graded response model and the generalized partial credit model. The graded response model was parameterized as:

$$P_{ij}^{+}(\theta) = \frac{e^{1.7a_i(\theta - b_{ij})}}{1 + e^{1.7a_i(\theta - b_{ij})}} \qquad (1)$$

where

$P_{ij}^{+}(\theta)$ is the probability of scoring/selecting category j or higher for item i, given trait score θ,

a_i is the item discrimination, and

b_{ij} is the category parameter (threshold) for category j in item i.

There is one less category parameter than the number of categories (the probability of choosing the first category or higher is one, so the first threshold occurs between the first and second categories, or scores 0 and 1) and a five-category item would have four category parameters. In PARSCALE, the category parameters are separated into an item location (constant for all categories within an item) and a category distance from the item location; for this study, the estimated category parameter was subtracted from the item location to put the PARSCALE estimates in terms of equation (1). In MULTILOG, there is no 1.7 in the function, so the a-parameter estimate from MULTILOG was divided by 1.7 to make it comparable to equation (1). The 1.7 was included in the model here to make the scaling commensurable with familiar dichotomous models.

The generalized partial credit model (not the traditional partial credit model, but a generalization with varying item discriminations) was parameterized as:

$$P_{ij}(\theta) = \frac{e^{1.7a_i \sum_{k=0}^{j}(\theta - b_{ik})}}{\sum_{j=0}^{m_i - 1} e^{1.7a_i \sum_{k=0}^{j}(\theta - b_{ik})}} \qquad (2)$$

where

$P_{ij}(\theta)$ is the probability of scoring/selecting category j in item i, given trait score θ (unlike the graded response model, it is the probability of scoring exactly j, not j or higher),

a_i is the item discrimination,

b_{ij} is the category parameter (step difficulty) for category j (the transition where j - 1 and j are equally likely), and

m_i is the number of categories for item i (numbered from 0 to m_i - 1 here).

As in the graded response model, there is one less step difficulty than the number of categories; there is no parameter for the first category because the first transition is between the first and second categories (0 and 1), and $\theta - b_{i0}$ is defined to be 0 for any θ (or the summations can start at j = 1 if one is added to the denominator). In PARSCALE, as for the graded response model, the category parameters are separated into an item location (constant for all categories within an item) and a category distance from the item location; for this study, the estimated category parameter was subtracted from the item location to put the PARSCALE estimates in terms of equation (2). In MULTILOG, the generalized partial credit model is obtained by putting constraints on the nominal response model (polynominal contrasts on the *a* parameters, with the quadratic and higher terms fixed to zero, and triangle contrasts on the *c* parameters). Childs and Chen (1999) described how the parameters and contrast matrices from MULTILOG can be transformed to the discrimination and step parameters in equation 2; in addition, for the present study the discriminations were re-scaled to take into account the constant of 1.7 instead of 1.

For each of these models (graded response and generalized partial credit), item parameters were simulated for a 10-item test[1], with 5 response categories in each item. Ten items would seem short for a dichotomous test, but would be realistic for a test with complex constructed responses, or an attitude survey. The logs of the discrimination parameters were randomly selected from a normal distribution with a mean of -0.5 and standard deviation of 0.2 (the discriminations themselves had a mean of 0.62 and standard deviation of 0.12; polytomous items can have relatively low discriminations compared to dichotomous items while still providing more information because each category adds to the item information). The first category parameters for each item was drawn from a uniform distribution between -2 and 1, and successive category parameters in the same item were 0.33 units apart. Different item parameters were used for each replication, because with the small number of items used for each test, idiosyncrasies in the particular set of items chosen (such as easy items being paired with low discriminations by chance) could have influenced the results if the same items has been used across replications.

Simulees were drawn from one of three distributions: normal [0, 1], uniform [-1.73, 1.73], or beta [2, 5.5], which produced a positively skewed distribution. The normal and uniform distributions had a mean of zero and a standard deviation of one, and the skewed distribution was rescaled (by subtracting 0.267 and multiplying by 6.59) so that it also had mean zero and standard deviation one. Both PARSCALE and MULTILOG can scale the item parameters such that the estimated latent distribution (the posterior quadrature distribution) has mean zero and standard deviation one, so there was no need for re-scaling and equating errors would not be compounded with estimation error. Each population distribution was crossed with two sample sizes: 250 and 500. Five hundred has been recommended as a minimum sample size for the

[1] Initially, there were plans to try a longer test as well to see if the two packages gave more similar results with a

graded response model (Ankenmann & Stone, 1992; Reise & Yu, 1990), and 250 was chosen as well to test whether differences between the packages were greater with smaller-than-recommended sample sizes.

One hundred replications were conducted with different item parameters and different sets of simulees. Because different item parameters were used in each replication, bias and RMSE were calculated across items as well as replications (for example, item 1 was different in each replication, so it was not particularly meaningful to calculate the bias for item 1 separately from the other items).

Calibration

In PARSCALE, the logistic metric was used with a constant of 1.7. The options FREE=(0,1) and POSTERIOR were used to estimate the posterior distribution after each E and M step and scale it to have a mean of 0 and a standard deviation of 1. Up to 100 EM cycles and 2 Newton cycles were allowed, with a stopping criterion of 0.01. Defaults were used for all other specifications. In the case of the partial credit model, a number of replications[2] either failed to converge or caused floating point errors or resulted in one or more items with extreme category parameters (absolute values greater than 5, generally in the double-digits). Most of these cases ran fine when prior distributions were used for the item parameters or when the constant was changed from 1.7 to 1 (discrimination parameters were later re-scaled to compensate), and a few needed both these changes and extra iterations.

In MULTILOG, 30 quadrature points, evenly spaced from -4 to 4, were used to correspond to PARSCALE's defaults. Up to 100 cycles were also allowed. Otherwise, default

---

longer test, but this seemed unnecessary in light of the results with the short 10-item test.

[2] For the samples of 500 simulees, these problems occurred in 18 replications for the normal distribution, 24 for the skewed, and 18 for the uniform. For the samples of 250 simulees, these problems occurred in 19 replications for the normal distribution, 19 for the skewed, and 10 for the uniform.

values were used for other specifications. The parameters were modified as described in the explanation of equations (1) and (2).

In both packages, trait parameters were estimated by direct maximum likelihood, not Bayesian methods (the Bayesian methods available are somewhat different in the two packages: MULTILOG uses modal-a-posterior estimation and PARSCALE uses expected-a-posterior estimation). Simulees with zero or perfect scores were omitted from the comparisons of theta estimates.

Analyses

The accuracy in parameter recovery was measured by bias and root mean square error (RMSE). Bias for an item or trait parameter was defined as the mean difference, across replications and items/people, between the estimated value and the true value.

$$\text{bias}_{\Lambda} = \frac{\sum_{j=1}^{n}\sum_{i=1}^{m}(\hat{\Lambda}_{ij} - \Lambda_{ij})}{nm}, \qquad (3)$$

where

Λ is an item parameter (discrimination or category parameter) or trait parameter,

Λ_{ij} is the i^{th} specific instance of Λ in replication j

$\hat{\Lambda}_{ij}$ is the estimate of parameter Λ_{ij} for replication j,

m is the number of instances of Λ in replication j (10 for the discrimination, 40 for the category parameters, 500 for the trait parameter), and

n is the number of replications.

RMSE was the square root of the average squared difference between the true and estimated values.

$$\mathrm{RMSE}_{\Lambda} = \sqrt{\frac{\sum_{j=1}^{n}\sum_{i=1}^{m}(\hat{\Lambda}_{ij} - \Lambda_{ij})^2}{nm}}, \qquad (4)$$

where the symbols are as defined for (3).

The estimates from of MULTILOG and PARSCALE were also compared to each other. If there were chance differences between the data samples and the population, or anything about the general marginal maximum likelihood procedure that would tend to produce inaccuracies, the estimates from the two packages would be more similar to each other than to the true values. The square root of the average squared difference between the estimates will be termed the root mean square difference (RMSD, similar to RMSE except that there are two estimates instead of an estimate and a true value for the parameter) and was calculated as

$$\mathrm{RMSD}_{\Lambda} = \sqrt{\frac{\sum_{j=1}^{n}\sum_{i=1}^{m}(\hat{\Lambda}_{ij} - \hat{\Lambda}'_{ij})^2}{nm}}, \qquad (5)$$

where

Λ, m, and n are as defined for (3),

$\hat{\Lambda}_{ij}$ is the estimate from MULTILOG of the i^{th} specific instance of Λ in replication j,

and $\hat{\Lambda}'_{ij}$ is the estimate of the same instance of the parameter from PARSCALE.

Results

The bias and RMSE and the difference and RMSD between MULTILOG and PARSCALE for each condition are reported in Table 1 for the discrimination parameter. The same information is shown in Table 2 for the category (step or threshold) parameters.

insert Table 1 about here

The discrimination parameters showed very little bias, though in a relative sense bias was greater for PARSCALE than for MULTILOG except when the trait parameters were uniformly distributed. RMSEs were similar across conditions, except that the MULTILOG partial credit RMSEs tended to be the smallest. The sample of 250 simulees also had negligible bias, with RMSEs about 45% (range 36%-46%) higher than in the sample of 500 for PARSCALE, about 50% (range 43%-56%) higher for MULTILOG graded response, and variable (range -6% to 37%) for MULTILOG partial credit. The RMSD between MULTILOG and PARSCALE remained small for the sample of 250, though it was higher than it had been for the sample of 500 for the graded response model (range 49% to 144%) and lower for the partial credit model (range -14% to -68%). The RMSD were small to begin with, so large percentage changes should be interpreted accordingly.

To obtain a more quantitative comparison of the factors, the variance in the logs of the absolute differences between true and estimated parameters was partitioned, using maximum likelihood methods available in the VARCOMP procedure in SAS 8.01. Because the items were different in each replication, it was not possible to calculate a RMSE across replications for each item, and the RMSE of an item parameter within a replication is simply the absolute value of the difference between the true and estimated values--because these values were highly skewed, the natural log transformation was used for the variance decomposition. The factors were software package, trait distribution, and sample size; replication and item within replication were left in the error term. The graded response and partial credit models were analyzed separately. Sample size accounted for 3% of the variance in the graded model, and 2% in the partial credit model. No other factor accounted for as much as 1% of the variance in the partial credit model, but the three-way interaction between package, distribution, and sample size accounted for 35% of the

variance in the graded response model. The three-way interaction was not due to any particular cell being unexpectedly large, and at this point it could conceptually be considered random error.

insert Table 2 about here

As seen in Table 2, there was virtually no bias in the category parameter estimates (the same was true for the sample of 250). MULTILOG and PARSCALE had nearly identical RMSEs. For the sample of 250, RMSEs were about 50% higher (range 42% - 67%), and the RMSD between MULTILOG and PARSCALE remained small with a tendency to be larger (range -17% to 82%, again percentages seem large because the base RMSD was small to begin with) than in the sample of 500.

For the category parameters, the RMSE was also calculated separately for each category. In Figures 1 and 2, the RMSEs are plotted for each category separately. The RMSE was slightly, but consistently, larger for the first and last category parameters, similar to the findings of Reise and Yu (1990).

insert Figures 1 and 2 about here

Again, the variance in the logs of the absolute differences between true and estimated parameters was decomposed, this time into variance due to software package, distribution, sample size, and item category. Sample size accounted for 2% of the variance under both the partial credit and graded response models, and the three-way interaction between package, distribution, and sample size accounted for 20% of the variance in the graded response model. This interaction appeared to be primarily due to larger RMSE for MULTILOG when the trait distribution was skewed, but only for the smaller sample of 250.

For the trait parameters, bias and RMSE/RMSD for the sample size of 500 are displayed in Table 3. There was essentially no bias in any condition. The RMSE for the samples of 250

(not displayed) were only 1% to 3% larger; this is consistent with other studies showing sample size has little effect on recovery of trait parameters, even when sample size impacts the accuracy of the item parameter estimates on which the trait estimates are based (Ankenmann & Stone, 1992; Choi, Cook, & Dodd, 1997; Reise & Yu, 1990; Stone, 1992). The RMSE was about 70-80% larger for the graded response model than for the generalized partial credit model. RMSE was nearly the same for MULTILOG and PARSCALE (and the RMSD between them was very small). RMSE did not appear to depend on the population distribution. Again, the trait parameters were estimated by direct maximum likelihood, not Bayesian procedures, so the parameters would have been influenced by the population distribution ***only*** through any effects of the distribution on the estimation of item parameters.

insert Table 3 about here

Bias is plotted by trait level in Figures 3-5. Simulees were grouped by theta into the following intervals: (<-3), [-3, -2), [-2, -1), [-1, 0), [0, 1), [1, 2), [2, 3), (>=3). Due to the way the uniform and skewed distributions were defined for this study, the uniform distribution had no simulees in the two upper or two lower groups, and the skewed distribution had no simulees in the two upper groups. When these extreme groups were present, however, the thetas within these groups were clearly biased towards the mean; very low thetas had estimates greater than the true thetas and very high thetas had estimates less than the true thetas. This appears odd because maximum likelihood estimates are typically slightly biased away from the mean (Wang, Hanson, & Lau, 1999;Wang & Wand, 2001). However, in this example no trait value was estimated for simulees who scored in the lowest category on all items or in the highest category on all items. Though these simulees were only a small proportion of the total sample, they were a sizable proportion of the groups with thetas < -2 or > 2 (40-50% in the two most extreme groups, around

16% (graded response) or 11% (partial credit) in the next most extreme intervals). The remaining simulees in these groups were those who scored higher (lower) than expected based on their thetas, thus the bias towards the mean. The bias was small in the remaining groups.

insert Figures 3-5 about here

Limitations

One limitation to the generalization of this study is that the category parameters were evenly spaced on the theta scale. Also, even the highest and lowest categories were never extreme relative to the simulees. This meant that sparse categories were rare; for the partial credit model the smallest categories averaged around 70 simulees and for the graded response model the smallest categories averaged just over 30 simulees. In data sets where sparse categories were more frequent, RMSEs would tend to be higher, at least for some parameters. However, there is no reason to suggest that the RMSD between MULTILOG and PARSCALE would be systematically different.

Conclusions

MULTILOG and PARSCALE item and trait parameter estimates were very similar, as indicated by the root mean square difference between them. This was true for both the graded response model and the generalized partial credit model, and for normal, skewed, and uniform trait distributions. Users can feel free to choose between MULTILOG and PARSCALE based on other factors, such as availability, ease of use, speed, and other personal preferences rather than accuracy.

References

Andrich, D. (1978). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement, 2*, 581-594.

Ankenmann, R. D., & Stone, C. A. (1992, April). *A Monte Carlo study of marginal maximum likelihood parameter estimates for the graded model*. Paper presented at the annual meeting of the National Council for Measurement in Education, San Francisco, CA. (ERIC Document Reproduction Service No. ED347189)

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29-51.

Carlson, R. D., & Locklin, R. H. (1995). *Item response theory: Comparing BILOG and MicroCAT calibration for a mathematics test*. Statesboro: Georgia Southern University. (ERIC Document Reproduction Service No. ED393881)

Childs, R. A., & Chen, W.-H. (1999). Software note: Obtaining comparable item parameter estimates in MULTILOG and PARSCALE for two polytomous IRT models. *Applied Psychological Measurement, 21*, 89-90.

Choi, S. W., Cook, K. F., & Dodd, B. G. (1997). Parameter recovery for the partial credit model using MULTILOG. *Journal of Outcome Measurement, 1*, 114-142..

Drasgow, F. (1989). An evaluation of marginal maximum likelihood estimation for the two-parameter logistic model. *Applied Psychological Measurement, 13*, 77-90.

Kirisci, L. Hsu, T. C., & Yu, L. (2001). Robustness of item parameter estimation programs to assumptions of unidimensionality and normality. *Applied Psychological Measurement, 25*, 146-162.

Masters G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Mislevy, R. J., & Stocking, M. L. (1989). A consumer's guide to LOGIST and BILOG. *Applied Psychological Measurement, 13*, 57-75.

Muraki, E., & Bock, R. D. (1997). PARSCALE: IRT item analysis and test scoring for rating-scale data [Computer software]. Chicago: Scientific Software.

Ree, M. J. (1979). Estimating item characteristic curves. *Applied Psychological Measurement, 3*, 371-385.

Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement, 27*, 133-144.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplements, 17*.

Seong, T. J. (1990). Sensitivity of marginal maximum likelihood estimation of item and ability parameters to the characteristics of the prior ability distributions. *Applied Psychological Measurement, 14*, 299-311.

Stone, C. A. (1992). Recovery of marginal maximum likelihood estimates in the two-parameter logistic response model: An evaluation of MULTILOG. *Applied Psychological Measurement, 16*, 1-16.

Swaminathan, H., & Gifford, J. A. (1983). Estimation of parameters in the three-parameter latent trait model. In D. J. Weiss (Ed.), *New Horizons in Testing: Latent trait test theory and computerized adaptive testing* (pp. 13-30). New York: Academic Press.

Thissen, D. (1991). MULTILOG: Multiple, categorical item analysis and test scoring using item response theory [Computer software]. Chicago: Scientific Software.

Thissen, D. J., & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika, 49*, 501-519.

Wang, T., Hanson, B. A., & Lau, C.-M. A. (1999). Reducing bias in CAT trait estimation: A comparison of approaches. *Applied Psychological Measurement, 23*, 363-278.

Wang, S. & Wang, T. (2001). Precision of Warm's weighted likelihood estimates for a polytomous model in computerized adaptive testing. *Applied Psychological Measurement, 25*, 317-331.

Yen, W. M. (1987). A comparison of the efficiency and accuracy of BILOG and LOGIST. *Psychometrika, 52*, 275-291.

Table 1

Bias, RMSE, and RMSD for Discrimination Parameters

	MULTILOG		PARSCALE		MULTILOG - PARSCALE	
Population Distribution	Bias	RMSE	Bias	RMSE	Difference	RMSD
			Graded Response, N = 500			
Normal	0.009	0.095	0.018	0.099	-0.009	0.019
Skewed	0.005	0.097	0.016	0.097	-0.011	0.038
Uniform	0.031	0.093	0.025	0.088	0.006	0.015
			Partial-Credit, N = 500			
Normal	0.010	0.072	0.030	0.121	-0.020	0.087
Skewed	0.005	0.079	0.020	0.085	-0.015	0.052
Uniform	0.031	0.076	0.027	0.094	0.004	0.064
			Graded Response, N = 250			
Normal	0.016	0.137	0.032	0.150	-0.016	0.037
Skewed	0.010	0.133	0.027	0.139	-0.016	0.057
Uniform	0.040	0.136	0.037	0.138	0.003	0.037
			Partial-Credit, N = 250			
Normal	0.019	0.105	0.033	0.114	-0.014	0.028
Skewed	0.014	0.115	0.031	0.116	-0.018	0.045
Uniform	0.038	0.110	0.028	0.103	0.010	0.025

Table 2

Bias, RMSE, and RMSD for Category (Step or Threshold) Parameters

	MULTILOG		PARSCALE		MULTILOG - PARSCALE	
Population Distribution	Bias	RMSE	Bias	RMSE	Difference	RMSD
			Graded Response, N = 500			
Normal	0.001	0.170	0.003	0.168	-0.002	0.044
Skewed	-0.017	0.178	-0.007	0.169	-0.010	0.045
Uniform	-0.002	0.148	-0.002	0.145	-0.001	0.017
			Partial-Credit, N = 500			
Normal	0.002	0.192	0.002	0.192	0.000	0.022
Skewed	-0.015	0.210	-0.002	0.195	-0.012	0.072
Uniform	0.000	0.198	0.000	0.191	0.000	0.044
			Graded Response, N = 250			
Normal	0.009	0.253	0.010	0.250	-0.001	0.036
Skewed	-0.020	0.297	-0.011	0.267	-0.009	0.083
Uniform	0.003	0.221	0.003	0.215	0.000	0.028
			Partial-Credit, N = 250			
Normal	0.011	0.278	0.011	0.277	0.000	0.036
Skewed	-0.012	0.297	-0.002	0.283	-0.010	0.092
Uniform	-0.003	0.283	-0.003	0.277	0.000	0.047

Table 3

Bias, RMSE, and RMSD for Trait Parameters

	MULTILOG		PARSCALE		MULTILOG - PARSCALE	
Population Distribution	Bias	RMSE	Bias	RMSE	Difference	RMSD
			Graded Response, N = 500			
Normal	0.001	0.671	0.001	0.675	-0.001	0.030
Skewed	0.002	0.679	-0.006	0.677	0.010	0.048
Uniform	0.000	0.650	-0.001	0.668	0.000	0.021
			Partial-Credit, N = 500			
Normal	0.002	0.374	0.002	0.377	0.000	0.018
Skewed	0.012	0.378	0.002	0.374	0.011	0.044
Uniform	-0.001	0.391	-0.001	0.377	0.000	0.150
			Graded Response, N = 250			
Normal	0.006	0.680	0.006	0.667	0.000	0.046
Skewed	0.003	0.695	-0.007	0.678	0.010	0.065
Uniform	0.001	0.653	0.001	0.657	0.000	0.032
			Partial-Credit, N = 250			
Normal	0.007	0.379	0.007	0.375	0.000	0.033
Skewed	0.016	0.389	0.003	0.378	0.012	0.052
Uniform	-0.001	0.372	-0.001	0.375	-0.001	0.032

Figure Captions

Figure 1. RMSE of category parameters, by item category and trait distribution, for the graded response model.

Figure 2. RMSE of category parameters, by item category and trait distribution, for the generalized partial credit model.

Figure3. Bias of trait parameters, by trait level, for the normal trait distribution.

Figure4. Bias of trait parameters, by trait level, for the skewed trait distribution.

Figure5. Bias of trait parameters, by trait level, for the uniform trait distribution.

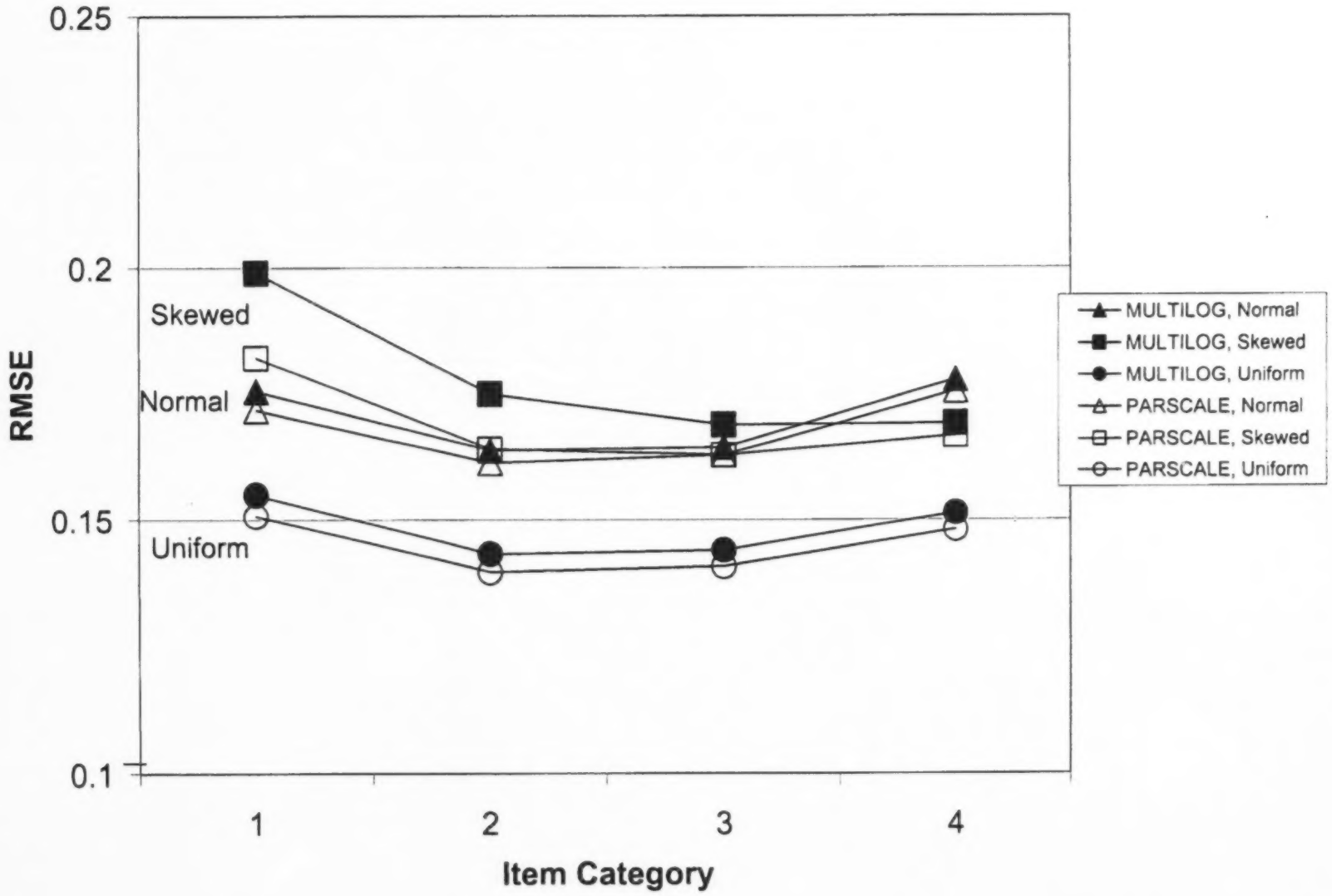

Figure 1. RMSE of category parameters, by item category and trait distribution, for the graded response model.

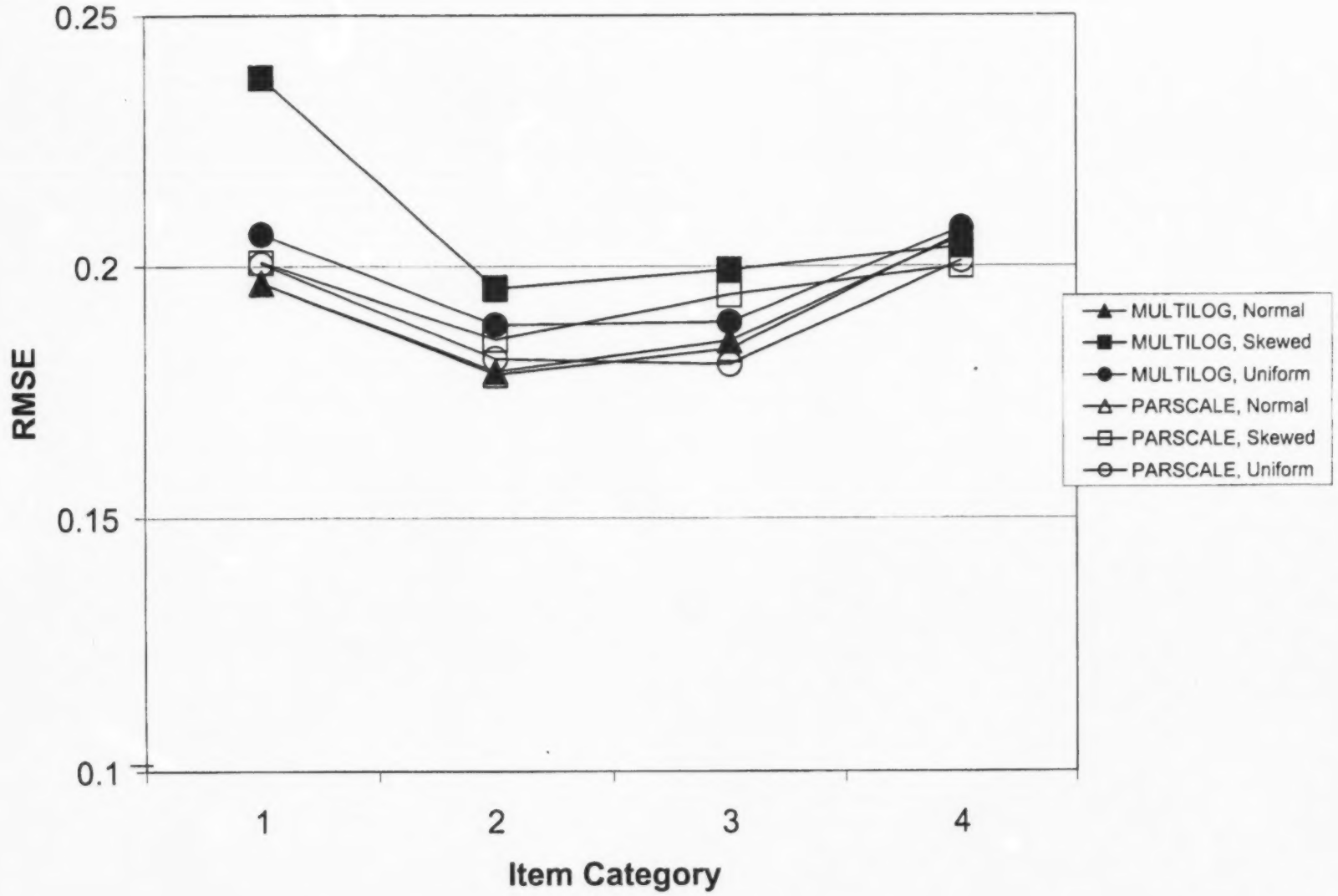

Figure 2. RMSE of category parameters, by item category and trait distribution, for the generalized partial credit model.

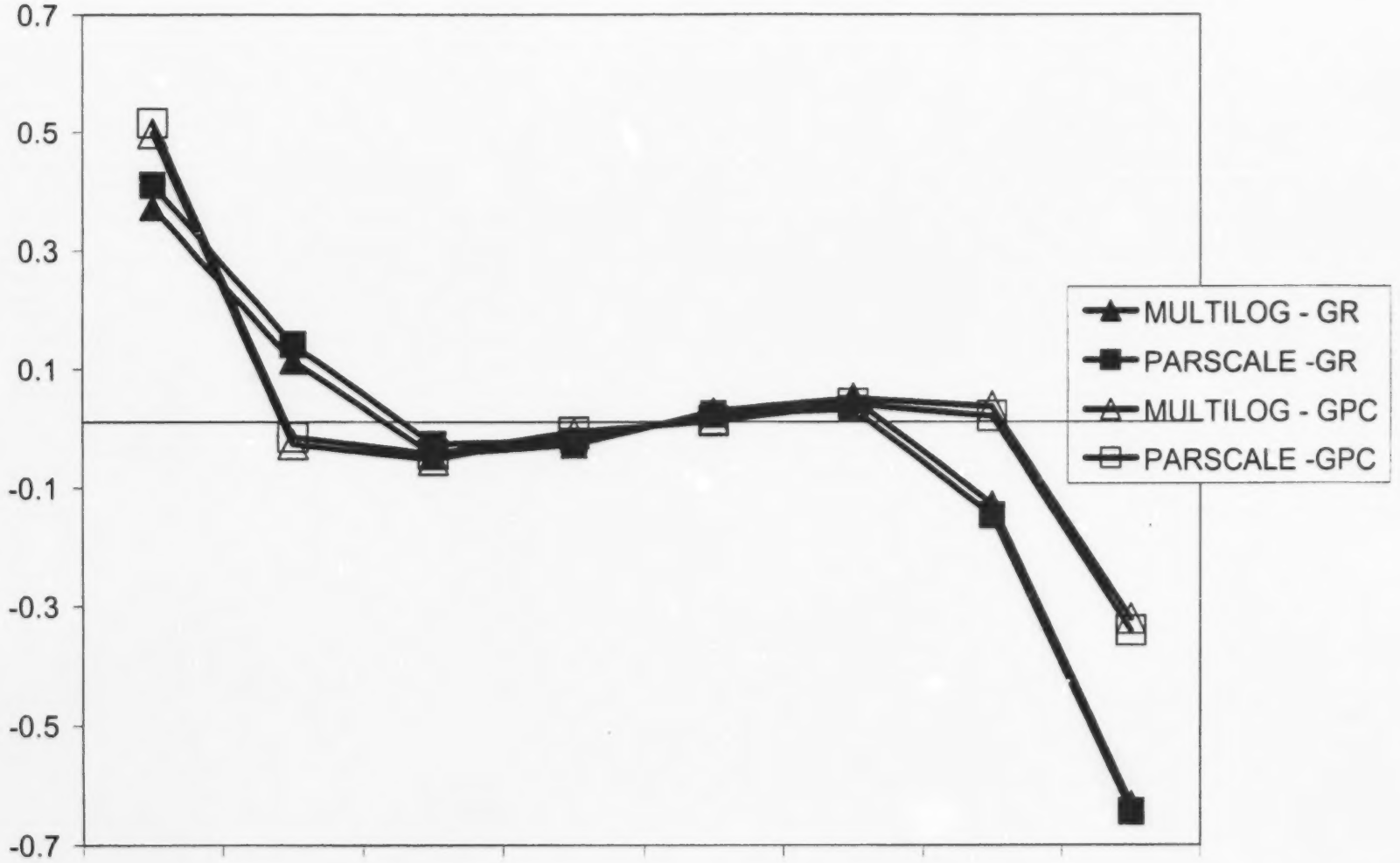

Figure3. Bias of trait parameters, by trait level, for the normal trait distribution.

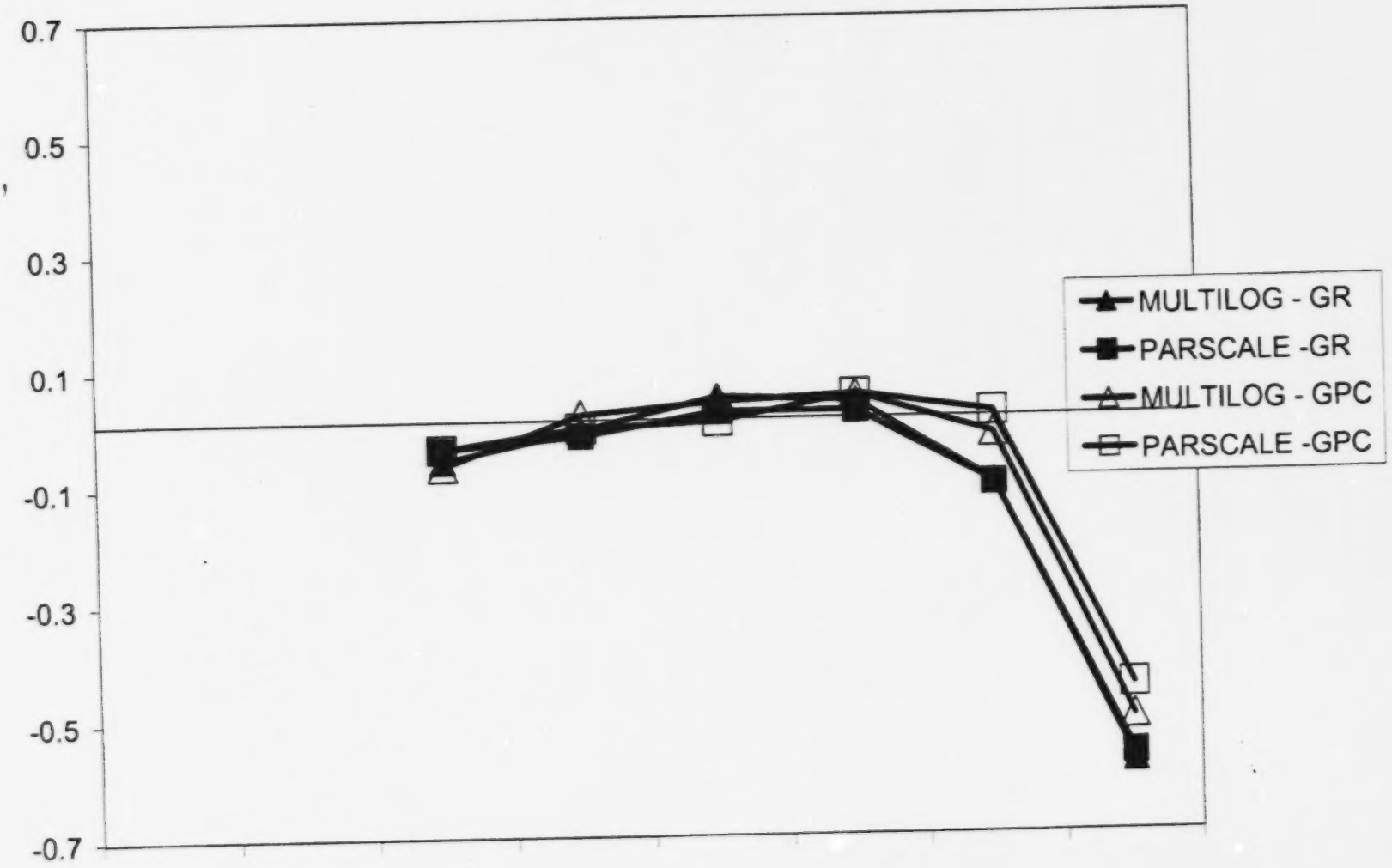

Figure4. Bias of trait parameters, by trait level, for the skewed trait distribution.

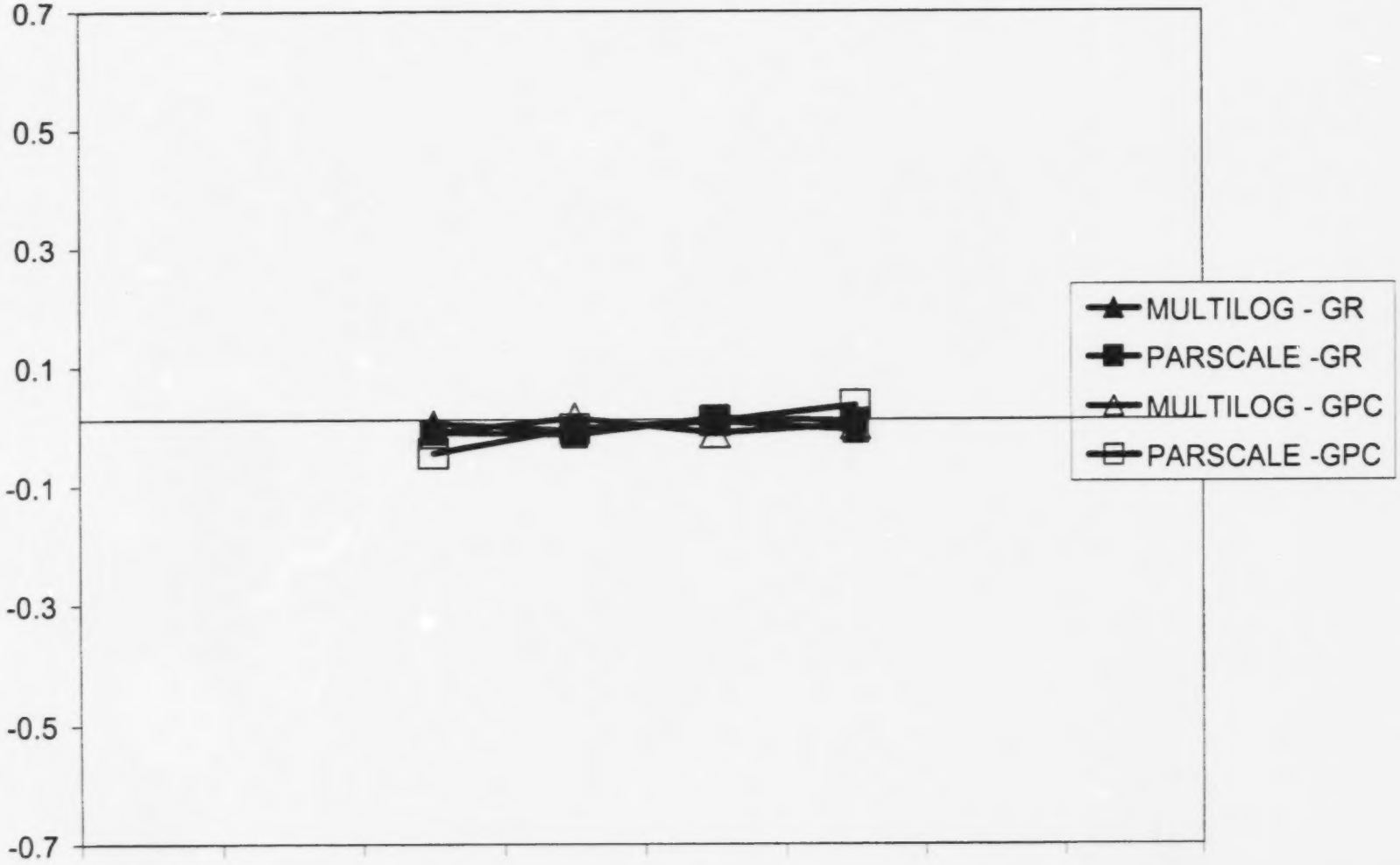

Figure5. Bias of trait parameters, by trait level, for the uniform trait distribution.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

TM034890

I. DOCUMENT IDENTIFICATION:

Title: Recovery of Graded Response and Partial Credit Parameters in MULTILOG and PARSCALE	
Author(s): Christine E. DeMars	
Corporate Source:	Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☑

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) and paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

☐

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

☐

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Signature: Christine E DeMars	Printed Name/Position/Title: Christine DeMars/assistant prof.
Organization/Address: MSC 6806, Center for Assessment & Research Studies James Madison University Harrisonburg, VA 22807	Telephone: (540) 568-8047 FAX: (540) 568-7878
	E-Mail Address: demarsce@jmu.edu Date: 4/2/03